Aalto University
School of Business

MULTI-TOUCH ATTRIBUTION IN THE MOBILE GAMING INDUSTRY

Master's Thesis
Joonas Syrjänen
Aalto University School of Business
Information and Service Management
Spring 2019

Author Joonas Syrjänen

Title of thesis MULTI-TOUCH ATTRIBUTION IN THE MOBILE GAMING INDUSTRY

Degree Master of Science in Economics and Business Administration

Degree programme Information and Service Management

Thesis advisors Pekka Malo, Timo Kuosmanen

Year of approval 2019 **Number of pages** 58 **Language** English

Abstract

User acquisition spend is a big investment for mobile gaming companies. Because of the large scale, even small improvements in how this spend is allocated can provide big returns. To allocate advertising spend well; it is important that the credit of a conversion be attributed as accurately as possible. The current attribution model - standard to the industry - is a last-touch attribution model, which attributes 100% of the credit to the last touch-point. However, before a user installs a game they might see ads from multiple channels that might all affect the user's propensity to install. With the last-touch attribution model, the uplift of these ads is not observed which skews the returns on advertising spent for different channels.

This study looks at how install probability develops as impressions per user increase, how long the effect of an ad lasts and attempts to find better attribution models that attribute credit better than the last-touch model. Three multi-touch attribution models are proposed; two based on the Shapley value and one based on the ad effect time decay of different channels. The data for this study comes from a mobile gaming company and consists of impressions seen by both installed and non-installed users as well as impression channels, impression time and install time. The data was collected during a 38-day period and has data from 44,719,217 users who were divided into a training set and a test set with a 70%/30% split. The test set is used to validate the proposed models against the last-touch attribution model by using the models trained on the training set to generate predictions on install probability for user paths in the test data set.

The study finds that the ad effect of all channels declines very quickly after the first day and is almost zero at seven days after the impression. The study also attempts to find the correlation between install probability and the amount of impressions a user has seen. Regarding this objective, the study is inconclusive. This correlation behaves very differently between different channels and because the amount of impressions per users could not be controlled for, it is difficult to deduce causation. Out of the three proposed attribution models, only one is able to outperform the last-touch model when it comes to predicting install probabilities from the training set's paths. The model that outperformed is a Shapley value based model that considers the times of impressions for each path when calculating credit attribution.

Finally, the study finds that only 9.5% of observed installs had impressions from more than one channel during a seven-day attribution window. This combined with the difficulty of validating attribution models based on return on advertising spend means that developing a multi-touch attribution model probably is not a very low hanging fruit for performance marketers. What would be worth looking into would be to test optimizing the frequency of ads shown to users.

Keywords Marketing attribution, performance marketing, Multi-touch attribution

Tekijä Joonas Syrjänen

Työn nimi Moni-kosketus-atribuutio mobiilipeli-alalla

Tutkinto Kauppatieteiden maisteri

Koulutusohjelma Information and Service Management

Työn ohjaajat Pekka Malo, Timo Kuosmanen

Hyväksymisvuosi 2019 **Sivumäärä** 58 **Kieli** English

Tiivistelmä

Käyttäjähankintaan laitettu raha on iso investointi mobiilipeliyrityksille. Investoinnin suuresta mittakaavasta johtuen pienetkin parannukset investointien allokaatioon voi tuottaa suuria voittoja. Jotta investointeja voidaan allokoida tehokkaasti, on tärkeää, että ansio käyttäjien hankinnasta atribuoidaan mahdollisimman tarkasti oikeille kanaville. Tällä hetkellä alan standardi on viimeisen kosketuksen atribuutio malli joka atribuoi 100% asennuksen ansiosta viimeiselle kosketuspisteelle. Joskus asiakas on saattanut kuitenkin nähdä mainoksen useammalta kuin yhdeltä kanavalta, jolloin hänen päätökseensä asentaa peli on voinut vaikuttaa useampi kuin yksi kanava. Tällaisessa tilanteessa viimeisen kosketuksen malli ei atribuoi muille kosketuspisteille ollenkaan ansiota mikä puolestaan vääristää kanavien tuomaa tuottoa.

Tämä tutkimus pyrkii selvittämään, miten asennustodennäköisyys kehittyy, kun yhden kuluttajan näkemien mainosten määrä kasvaa, kuinka nopeasti mainoksen vaikutus nollaantuu sekä löytää parempia atribuutio malleja. Tutkimuksessa esitetään kolme vaihtoehtoista mallia: kaksi Shapley-arvosta johdettua mallia ja yksi mainoksen vaikutuksen nollaantumisaikaan perustuva malli. Tutkimuksen data tulee mobiilipeliyritykseltä ja sisältää asentaneiden ja asentamattomien kuluttajien impressiot, impressioiden ajat, impressioiden kanavat sekä asennusajat. Data kerättiin 38 päivän ajan ja sisältää 44,719,217 käyttäjää jotka ovat jaettu koulutus- ja testauskohortteihin 70%/30% jaolla. Testikohorttia käytetään validoimaan mallit viimeisen kosketuksen mallia vastaan luomalla mallien ja koulutuskohortin perusteella ennusteet testikohortin käyttäjäpolkujen asennustodennäköisyyksistä.

Tutkimuksesta käy ilmi, että mainosten vaikutus vähenee huomattavasti ensimmäisen päivän jälkeen mainoksen näkemisestä ja on melkein nolla seitsemän päivän jälkeen. Tutkimus pyrkii myös löytämään korrelaation käyttäjän näkemien mainosten määrän ja asennustodennäköisyyden välillä. Tätä löydöstä tutkimus ei kuitenkaan pysty tuottamaan. Kanavat käyttäytyvät keskenään hyvin eri lailla kun katsotaan edellä mainittua korrelaatiota. Lisäksi käyttäjien näkemien mainosten määrää ei tutkimuksessa pystytä hallitsemaan mikä tekee korrelaation ja kausaation määrittelemisestä vaikeaa. Validoiduista malleista vain Shapley-arvoon perustuva malli - joka ottaa huomioon impressioajat - pärjäsi viimeisen kosketuksen attribuutiomallia vastaan.

Tutkimuksesta käy myös ilmi, että vain 9.5% asennuksista oli impressio useammalta kuin yhdeltä kanavalta seitsemän päivän atribuutioikkunan aikana. Kun tämän lisäksi otetaan huomioon se, että atribuutiomallien validointi markkinointi-investoinnin tuotolla on erittäin vaikeaa, ei moni-kosketus-atribuutiomallin kehitys ja jalkautus ole varmaan tuottavin investointi mitä yritys voi tehdä performanssimarkkinointitiimissään. Sitä vastoin mainosmäärän optimoinnin testaaminen olisi melko helppoa ja voisi olla hyödyllistä.

Avainsanat Markkinointi-atribuutio, Performanssimarkkinointi, Moni-kosketus-atribuutio

Table of Contents

List of Tables

List of Figures

1 Introduction

1.1 Motivation

The aim of this thesis is to find a multi-touch attribution model (MTA) that would be better than the industry standard last touch attribution (LTA) model. A better model would attribute the credit of a customer conversion more accurately to different touch-points in the user's conversion funnel. In other words, this thesis attempts to find a model that better estimates how much each touch-point in a user's conversion funnel contributed to the user's conversion. Different attribution models will be assessed by how well they can predict the conversion probabilities of different user paths drawn from a test dataset.

Mobile games are a huge and fast growing industry. NewZoo (2018), a games, esports and mobile market intelligence provider, forecasts that 2.3 billion users will spend $137.9 billion on games in 2018. Out of this $137.9 billion, mobile games are forecast to generate $70.3 billion with growth of 25.5% year on year. Mobile gaming companies use vast amounts of money to acquire new users and with user acquisition costs increasing this spend will only go up.

Paid user acquisition (UA) is often a necessity to achieve significant growth. As smart phones become more and more available growth in new users will start to slow. More and more businesses are also vying for user attention in the digital world. Increasing demand for impressions and a saturating supply of users means that cost per install (CPI) will go up as more and more businesses compete for a stagnating number of users. We are seeing this already: CPIs have increased in the past and continued to increase in 2018 (Olenski, 2017; Takahashi, 2014). With increasing user acquisition costs the return on ad spend will decrease if monetization stays the same. For companies depending heavily in UA this is a major challenge. The answer to increasing user acquisition costs is either increasing the lifetime revenue of a game's players or targeting potential users more effectively. By targeting potential users more effectively, companies can decrease the cost per install (CPI) increasing their return on ad spend – also known as ROAS. This can be done either by using ads that have a higher conversion rate or spending money on channels that convert users more effectively. Attribution models are used to credit different channels for the conversions they create and are essential to assessing the efficacy of marketing channels.

There has been earlier research on modelling multi-touch attribution (Dalessandro et. al., 2012; Yadagiri et. al., 2015; Shao and Li, 2011; Abishek et. al., 2012) but to my knowledge, this research has not been used in the mobile gaming context. An interesting feature of mobile games is that most of them are free to play. By offering the product for free, people are a lot likelier to try it (Lowell, 2013). With a price point of zero people are a lot more likely to try a product than one with a price, no matter how small the price is. This is because removing the price also removes the need to make a decision, which in itself takes effort. Previous studies on MTA (Kannan et. al., 2016) have not been made with data from the mobile gaming industry. They have focused more on industries with more traditional business logic. In addition, the majority of the articles (e.g. Shao and Li, 2011; Abishek et. al., 2012) that I have reviewed, have featured search as a major channel – this research does not. Because the threshold to try new mobile games is inherently very low, the importance of search may be smaller and the last touch-point may have a lot larger contribution to a conversion than in other industries. As such, MTA might give a more similar output as the LTA model in mobile gamin compared to other industries. A In addition, this research will make use of a very large data set that can further validate the findings of earlier research.

1.2 Research objectives

This thesis will attempt to find an MTA model more accurate at attributing credit to a user path's touch-points than the LTA model. Because the effects of individual touch-points cannot be observed, the models will be validated based on how well they can predict the install probabilities of different user paths in the test dataset. In order to determine in what time-window credit should be attributed, I will look at how the effect of an ad decays over time. I will also look at how the install probabilities of users develop as the amount of impressions they see increases. Lastly, I will look at how many installs would benefit from an MTA model.

The theoretical part of this thesis will first look into the underlying process of how people make purchase decisions to understand the reality that attribution models try to reflect. The major goal of the theoretical part is to evaluate different attribution models and help choose an MTA model to test. Finally, the thesis will attempt to create a validation method that allows comparing the proposed models against the LTA model.

The main research question of this thesis is:

How well would a multi-touch attribution model attribute conversion credit compared to the last-touch attribution model?

In addition, the research will attempt to answer the below questions.

1. What share of installed users have a touch-point from more than one channel?
2. How does install probability develop as impressions per user increase?
3. How does the effectiveness of ads decay overtime?

1.3 Structure

The second chapter of this thesis a look into previous research on purchase decision making, the mobile advertising ecosystem and attribution modelling. In the third chapter, the dataset used in this study will be explained and findings on ad effect time decay presented. In the next chapter, I will discuss and explain the models to be tested. In chapter five, the validation model will be explained and in the sixth, the research questions will be answered, and the results discussed. In the final chapter, limitations and future research avenues are expressed along with managerial implications.

2 Literature review

2.1 Consumer decision process

Before diving deeper into the details of the mobile ad system and different attribution models, I believe it is valuable to first look at how users end up making their purchase decisions. This is relevant because the goal of attribution models is understanding how different touch-points affect a user's decision-making process. This will help us understand better the reality MTA models try to reflect. For clarity, I will next define what I mean by user touch-points in this study.

I define user touch-points as any digital event where a user is exposed to the product. Touch-points can be thought of existing in a conversion funnel throughout which some users drop off and do not convert. In this research, the top of the funnel is defined as the first payed impression a user sees, and the bottom of the funnel is the user opening the game. There might be touch-points before bought engagements and there are touchpoints after opening the game e.g. re-engagement. However, because of the difficulty in gathering and combining consistent data from these touch-points, they will be omitted.

2.1.1 Factors affecting a consumer's decision-making process

Bettman et. al. (1991) approach a consumer's decision-making process as a choice task that is affected by the following factors: alternatives, attributes of value and uncertainties. This choice task becomes more difficult as the number of attributes and alternatives increases: the attribute values are difficult to gauge, the uncertainty about the values of different attributes is big and as the number of shared attributes between alternatives become smaller. Another factor in the difficulty of choice is the available information and its structure, this information is conveyed not only by ads and packages but also by price.

A consumer's decision process can also be approached from a model of consideration sets (Shocker et. al. 1991). In this model, a universal set of options exists that the user may or may not be aware of. Within this set is an awareness set consisting of information that the consumer knows either consciously and sub-consciously. This information exists in the long-term memory. A sub-set of the awareness set is the consideration set which has different alternatives

that the consumer chooses from. The decision itself is done in the consumer's working memory and different elements can be recalled or dropped fluidly in the decision process.

In the context of mobile gaming there are a huge number of alternatives and a lot of variety in the attributes of alternatives – though there are genres with very similar core mechanics e.g. Candy Crush's match-3 mechanic. On the other hand, the uncertainty i.e. risk – at least on the part of price - is minimal since the games are mostly free. I think that the consideration sets in the mobile gaming context depends on how users convert after interacting with a touch-point. First, it needs to be determined if more touch-points actually increases the probability of a conversion. If quantity leads to better probability, we can assume that the touch-points do affect the consumer's long-term memory. After verifying this, we need to look at where the increased probability is seen. When a user sees an ad, they can click the download button at the end of the ad that will take him to the application's store page. Here they can look at the details of the game and download it. A second scenario would be that the user closes the ad but downloads the game later in the application store.

If it was observed that, more touch-points lead to a higher click-through rate from ad to application store but not a higher click-to-install rate, we could assume that the touch-points only affect the phase where users are thinking of clicking the ad. In this phase, the consideration set would then be either to click the ad and stop what the user is currently doing or to carry on. In this scenario, it could be assumed that touch-points affect a user's impulses because the install decision would have been very much an impulse decision. On the other hand, if it was observed that more of touch-points lead to more searches for the game in the application stores, it could be assumed that both the awareness set, and the consideration set would be much larger. This is because there are many different choices to choose from the application store. With this scenario it could be assumes that the touch-points have a deeper affect because separately going to the application store, searching and downloading the app implies a higher intent than just clicking on an ad and downloading.

2.1.2 Planned versus impulse purchase decisions

Engel and Blackwell (1986) define an impulse purchase as: "*a buying action undertaken without a problem previously having been consciously recognized or a buying intention formed prior to entering the store.*" In mobile gaming, the definition of the application store depends on how a user ends up there. If a user clicks an ad, they will land straight to the advertised application's page. In a scenario like this, I would define the ad as an extension of the store as it is only a click and a few seconds away from the actual store. If a user opens the application store separately to the ad impression, I would not consider the ad a part of the store.

A study on impulse purchases in infomercials by Agee and Martin (2001) found that consumers that had seen an infomercial multiple times were more prone to make a planned purchase decision. In the study, impulse buyers had less previous interest in the product than the planned buyers did. Planned purchasers were more interested in demonstrations on product performance and testimonials. These findings point toward a positive correlation between the amount of impressions per user and install probability.

The phenomenon of ads affecting consumers over a long period after an impression is called the lag effect of advertising. The lag effect is affected by memory which in turn is conversely correlated with the process of learning (Bean, 1912). With repetition positively affecting learning (Cacioppo & Petty, 1979) it is easy to see why the amount of touch-points per user would also be positively correlated with conversion probability. Because memory is a factor in the lag effect of advertising, the effect of an individual touch-point decays as time goes by as people forget non-consequential information over time. However, the time decay that a touch-point experiences is not always the same and is affected by multiple factors. Berkowitz et. al. (2001) have shown this in their study on the lag effect of advertising. They found that different channels (newspaper and bill boards) have different lag effects.

So far, a few frameworks on how to approach the consumer purchase decision have been presented. Bettman et. al. think of the consumer purchase decision as a choice task where the consumer must choose between different options. Shocker et. al. think of the purchase decision more as a funnel – though they do not use this phrase- where the choice sets are consciously or unconsciously narrowed down. Due to the limitless nature of digital products I find the consideration sets of Shocker et. al. a more relevant framework because consumers could

essentially pick as many products as they want. However, the approach of consideration sets depends on how users convert. If they convert through planned purchase, they will go to the application store where the consideration set will include more options because search brings up other games. If on the other hand, users convert straight after seeing an ad, the consideration set will only include: continuing to play the current game, downloading the advertised game or stop playing.

With the above reasoning there two different type of conversions: an impulse decision and a planned decision. There is some evidence that the probability of both impulse and planned purchases are positively correlated with the amount of touch-points a user has interacted with. However, as Agee and Martin mentioned, the types of ads that are effective for these conversion types are different. It is also known that the reason the number of touch-points can be positively correlated is because repetition positively affects learning which in turn positively affects memory. This finding also gives a hint at ad effect time decay and that the decay speed differs between different channels.

2.2 Digital ad ecosystem & attribution windows

In the mobile ad ecosystem users are identified by IDs of their mobile phones. In general, when a user uses an app and there is an ad opportunity (e.g. user clicks to watch an ad to earn credits in a game) the app sends information of the opportunity – e.g. ID, format, source – via a supply side platform to a market platform. In the market platform there are integrate many demand side platforms (DSP) through which advertisers buy inventory. Advertisers can bid for the inventory either directly through DSPs or via video networks who then use their own DSPs. The buyers decide based on their information and monetization how much to bid on the ad opportunity. On the market platform the highest bid wins and the winner gets to upload their ad to the placement the user will see. All this takes a fraction of a second.

When a user has seen and/or clicked on an ad a post-back is sent to a mediation platform which logs the event for the video network - i.e. channel - that made the ad happen. If the user ends up downloading and opening the game another post-back is sent to the mediation platform that then attributes the install to a channel based on the attribution model in use. With the LTA model, the install would be attributed to the channel whose ad was the last the user saw – if it

happened within the model's attribution window. An attribution window is a span of time before an install during which a user must have had touch-points, for the install to be attributed to paid ads. Often it is defined that a click overrides an impression for a conversion because a click is seen as a stronger indicator of intent. To clarify: if a user sees an impression within an acceptable time window after the user has clicked on an ad the conversion is still attributed to the ad the user clicked. This is still called an LTA model even if the last impression did not get 100% of the credit.

When it comes to video networks, gaming companies usually pay per install. The networks however, pay for ad inventory by paying for a thousand impressions. After buying ad inventory, the networks sell the inventory to their customers based on how much they will be able to get for the thousand impressions. This is determined by how much the buyer is willing to pay for an install and how many installs the buyer is likely to get from the impressions. The amount of installs customer can generate from the impressions depends on how well the audience fits with the buyers product, how good the buyers ads are and how the buyer attributes installs. A malicious network could try to trick the LTA model by spamming very low-quality ads – e.g. banner ads – in order to maximize the probability that their ad would be the last ad a user saw before converting. These fraud cases can be very costly. For example, Uber sued its marketing partner for wasting tens of millions of dollars on ad fraud (CNBC, 2017).

2.3 Attribution models

Before converting, a user might often come into contact with multiple touch-points. This brings up a problem: how much did each of these touch-points affect the user's decision to install the game? How the credit for the app-open is shared among the paid touch-points is called attribution. Currently the default attribution model in the industry standard in performance marketing is last-touch attribution (Dalessandro et. al., 2012). This model attributes 100% of the credit from a conversion to the last touch-point. This attribution model implies the assumption that only the last touch-point had an effect on the user. Research done by McKinsey (2015) suggests that this is not accurate. Using a dataset of about 15000 households covering around 70% of discretionary spending, their research shows that brands with more digital touch-points were likelier to be selected by customers. This finding would suggest that customers are affected by multiple touch-points in the funnel.

A better way to attribute the credit would be to attribute it by how much different touch-points affect a user's decision. This is called multi-touch attribution (MTA). With a better attribution model mobile gaming companies could focus more advertising spend on channels that have a bigger effect on a user's probability to convert. This would increase the number of installed users. If other parts of the funnel stayed the same, focusing more on the most effective touch-points would mean more users would go through the funnel to opening the game. This in turn would decrease CPI and increasing ROAS.

There are many challenges with MTA. One is modelling; MTA models try to reflect a truth that cannot be observed. Even users themselves have at best only a vague idea on how different ads affected their decision. Thus, an MTA model can only estimate the effect of different touch-points. This makes it difficult to validate how well a model reflects reality. Furthermore, implementing an MTA model can be difficult. Because an MTA model divides credit amongst multiple marketing channels, it is important that all channels understand and accept the model. This also creates a risk that a channel might try to fraud the model to get more credit – though this also applies to LTA. An MTA model also requires a huge amount of data from each channel, as each individual click and impression from each channel needs to be stored. Currently the biggest channels Facebook and Google do not share impression and click data at the user level which limits the accuracy of and MTA model.

A major reason MTA is not used more often, is that it is difficult to implement and it can be difficult to justify the effort. To be able to justify the investment on an MTA system, marketing managers need to give an estimate on how much implementing it could improve the bottom line and to be able to get the different channels to comply they need to be persuaded to trust the accuracy and fairness of the model. Finally, validating attribution models in a sufficient way is difficult.

2.3.1 What makes a good attribution model?

Multiple different attribution models have been proposed to shed light on how different touch-points in the conversion funnel really affect a user's decisions. Dalessandro et. al. (2012) propose three subjective measurements to assess how good an attribution model is. These three measurements are:

" 1. Fairness - a good attribution system should reward an individual channel in accordance with its ability to affect the likelihood of conversion.

2. Data driven - A good attribution system should be derived specifically for the advertising campaign in question, using both ad treatment and conversion data captured during campaign

3. Interpretability - A good attribution system should be generally accepted by all parties with material interest in the system, on the basis of its statistical merit, as well as on the basis of intuitive understanding of the components of the system."

Interpretability is a factor that also other studies have brought up (eg. Shao & Li, 2011). It is for marketers to be able to interpret an attribution model's output and how it came about. This is important because marketers need to be able to quickly spot problems in the channels and explain their investments. Fairness and interpretability are also important in an attribution model because the model determines how to divide marketing budget between channels and thus determining the profit the different channels earn. It is imperative that all channels agree on the model's fairness and understand how it divides credit so that they can optimize their profit.

2.3.2 Carryover and spillover effects of advertising

An underlying assumption with MTA is that when a user interacts with a touch-point there are carryover and spillover effects. To understand the dynamics of these effects Kireyev et. al. (2016) use a vector error correction model (VECM) on aggregate data of display and search ads. A VECM allows estimating short- and long-term effects of different time series against one another. An error correction model estimates how fast a dependent variable returns to a state of equilibrium after other variables are changed. In the context of advertising, the variables could be e.g. number of impressions per ad format or channel or user action in the form of

clicks. In other words, the researchers use the model to estimate how long the effects of certain touch-point linger. The effect that ads have on user conversion over time is also called ad stock.

With VECM, this is estimated by looking at the dependent variables and the time it takes for them to return to normal after a change. For example, a campaign is put live that doubles the impressions of video ads for the weekend. After the campaign is live, we see an increase in the clicks on search ads. After the campaign is over, we still see a higher percentage of click in the first few days but with a declining trend after which the search clicks return to the level they were before the campaign. A VECM can more accurately estimate the effect the video campaign has on the search clicks. The researchers found that by using a VECM, the cost per acquisition (CPA) was lower than with a standard calculation of CPA and the ROAS higher than the ROAS calculated by the standard way. From their model, it is also easy to interpret the optimal budget allocation to the channels.

There are several other models to estimate what effect an advertisement has as time since last exposure increases. Naik (1999) defines the half-life of an advertisement as "*the time required for advertising effectiveness to wear out to one-half of the initial effectiveness level*." He estimates this time with by using a Kalman filter – also known as linear quadratic estimation (LQE). This is an algorithm that is often used in time series analysis. The algorithm has two steps. First, the Kalman filter creates estimates of the current variables with their uncertainties, then the outcome of the next measurement is received, and the estimates are updated with a weighted average, the estimates with more certainty being weighed more.

Using the Kalman filter on user level data related to an ad (number of impressions, time of individual impression, user conversion time) the lifetime of the effect of an ad can be estimated. The problem with using the Kalman filter in advertising scheduling is that it assumes an environment where there is only one advertiser and thus does not properly take into account how the advertising campaigns of other advertisers' effect ad half-life (Naik et. al., 1998). This is something that the VECM does take into account. The studies on the Kalman filter and VECM are limited in applicability to this thesis. Both studies approach the time decay of ad effectiveness on an aggregate level with the business outcomes focused on campaign planning rather than looking at how the time-decay of multiple touch-points on a user's funnel affect the probability of conversion. Never the less, the models these papers introduce do estimate the effect time has on a touch-points effectiveness and reflect how people forget information.

However, integrating time-decay to an MTA model that takes into account other aspects as well can be difficult.

2.3.3 Shapley value

In their research Dalessandro et. al. (2012) propose using Shapley value for attributing the credit of a conversion. This approach "assumes a simultaneous joint advertising treatment" which does not depict reality correctly as the treatments would not happen simultaneously but one after another. However, this simplification decreases the parameters needed to take into account. It also allows dividing joint treatment interaction outcomes equally.

The Shapley value used by Dalessandro et. al. in their attribution model was created by Lloyd Shapley in 1953 and estimates the value an individual player creates to a common goal in a collaborative transferrable utility (TU) game setting. In a collaborative game setting, players can make agreements on how to distribute the profits arising from coalitions the players make (Peters, 2008). In a collaborative transferrable utility game, it is assumed that the profits created by a coalition can be expressed with a single number. These characteristics apply well to the attribution problem where different touch-points are the players of the collaborative game. In the game all the touch-points contribute to the probability of a user converting and vie for the monetary payment created from it. With the help of Shapley values credit for achieving a common goal can be attributed to the players involved.

Shapley value is a value that calculates the payment to each individual player by creating a coalition of players one player at a time. Each player of the coalition demands their added value as payment. For each player the average of payments arising from different combinations the player can be in is the player's shapely value. In other words, the Shapley value assigns each player his or her average marginal contribution across the combinations. Because the basic Shapley value function has to go through all the different combinations players can be in the computations needed grow exponentially as the number of players grows. However, the amount of computations can be reduced by using the formula suggested by Shapley (1953), which does not calculate the marginal value of player for combinations that the player is not in.

$$\emptyset_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N|-|S|-1)!}{|N|!} \left(v(S \cup \{i\}) - v\,(S)\right), \quad (1)$$

where S is a coalition of players, $v(S \cup \{i\})$ is the worth of a coalition S without i, $v\,(S)$ is the worth of coalition S, i is a player and N is the number of players and the sum is all the coalitions with player i.

Function 1 forms coalitions one player at a time with each player's contribution being the coalition's value with the player minus the coalitions value without the player $v(S \cup \{i\}) - v(S)$. In the context of this study the players would be touch-points, the coalition a user path and the coalition without the latest player another user path without the latest player. The values of the coalitions would be the install probabilities of the paths.

A potential problem with using Shapley value for MTA is that it only takes into account the properties of the players in the collaborative game. In the context of mobile advertising it would not take into account what effect time-decay has on a touch-point - unless the information is attached to the player of the collaborative game. This would mean that each player would be a touch-point from a certain channel at a certain time. Depending on the granularity of time needed this could significantly increase the number of players and coalitions, increasing required computational capacity. A small granularity level of time could offer a better accuracy but would lead to a high amount of computation and small coalition sizes.

Another problem in giving players more details, is that the cohort size of users making up a coalition's data will decrease the more granular the player information gets. The smaller the cohort size is for each player the more volatility there will be with regards to the outcome. This could lead to some coalitions having very high conversion rates (probabilities) which in turn would lead to some players having very high Shapley values. An MTA model based on this kind of data could be very volatile in allocating credit to different touch-points. To sum up, the granularity of an attribution model depends on the amount of data and the amount of computational power available.

Using a model that derives from the Shapley value model on data from a real un-targeted campaign featuring seven channels, Dalessandro et. al. find that the two worst channels in terms of effectiveness benefit from a last-touch attribution model. However, they also find that using

a multi-touch attribution model does not largely change credit attribution overall from an LTA model if search-ads are not present. This could be because the channels are largely on the same level in the attribution model and do not push users to touch-points lower down the funnel – except to the search ads.

Berman (2018) also approaches attribution in a game setting but where information is symmetric among advertisers and publishers, as the previous researches assume an asymmetric information setting. Through his research Berman finds that through more sophisticated attribution models, the cost per conversion goes down. The research also looks at the how efficient the Shapley value is as an attribution model and attempts to take into account how CPIs and competitor actions change when an advertiser changes their bidding model. Surprisingly, by considering other players too, Berman finds situations where an LTA model leads to lower profits than not attributing credits at all. Yadagiri et. al. (2015) also use the Shapley value in their non-parametric model to divide "left-over return" to the different channels. The model first uses the Shapley value to attribute conversions, much in the same way as Dalessandro et. al. What is different in their model is that on top of the Shapley value they use Nash Bargaining to divide the "left-over return" to the different channels.

The probability of a player coalition (user path) that Shapley value needs is calculated by dividing all converted users of single coalition by all users of a coalition. This means that if there are no users with a certain coalition the probability of that coalition cannot be calculated. To estimate the probability of such player coalitions Yadagiri et. al. take all the coalitions without data and break those coalitions down to sub-coalitions that have users and average the probability of the sub-coalitions of each coalition without users. These averages are assigned as probabilities for the coalitions without data. The researchers also propose a semi-parametric model that uses a binary classification algorithm to estimate unknown coalitions. In this algorithm users are given a response variable of 1 (user converted) or 0 (user did not convert). Each user is also assigned a feature vector that has all the touch-points the user has been exposed to. To model the probability of an unknown coalition, the model can use either logistic regression or random forest.

After the probability for each coalition is calculated, the difference between baseline probability – i.e. organic probability - and the coalition's probability is calculated. This difference in probability is the value the coalition of players generate for the co-operative game. With this

value, the model uses Shapley value to divide credit between touch-points. The baseline probability is also attributed to the touch-points using Nash bargaining. Nash bargaining is a bargaining problem with an external option – in this case the benefit attributed by the Shapley value – that is reverted to if the touch-points do not co-operate. If the players do co-operate they get both the external option and the baseline probability attributed to them.

I think that the model Yadagiri et. al. propose improves upon the model Dalessandro et. al. propose. The most important improvement is the estimation of conversion probabilities of touch-point coalitions without user data. This makes the Shapley value model much applicable in many different scenarios. I did not find their argument for attributing the credit of organic conversion probability to touch-points very compelling. Cannibalization in user acquisition is when users who would have installed the game anyway are brought. This is seldom seen as a good thing. I think an attribution model that penalizes channels for cannibalizations would be much more beneficial. As with Dalessandro's model this model also suffers from not taking into account ad effect time-decay.

2.3.4 Bagged logistic regression

Where Dalessandro et. al. proposed a Shapley value based MTA-model Shao and Li (2011) propose using bootstrap aggregated logistic regression which gives the contribution of a touch-point as a coefficient estimate. Bootstrap aggregating - aka. Bagging - is a method where smaller training sets are drawn with replacement from the actual training set. After models are trained on all the sub-sets the outputs of these models are averaged. With the data used by this thesis, the training set would consist of user level impression and conversion data, where the conversion data would be the binary outcome and the touch-points (impressions) would be the coefficients that the model would assign the credit to. Each user would be a single individual in a sub-set and the credits attributed to different touch-points by each sub-set's model would be averaged across all the sub-sets.

Using bagging Shao and Li calculate the average misclassification and average standard deviations of the misclassification rates of different coefficients across the subsets created by the bagging method. They do this by comparing the credit of different touch-points (uplift in conversion probability) generated by each sub-set's model to an independent test data set, calculating the standard deviation of the misclassification rate of each of the coefficients and

averaging these deviations. By these averages, the researchers can estimate how good an MTA model is. Having a low misclassification rate means that the models are overall accurate in crediting the different touch-points and having a low average standard deviation means the model gives a stable estimate. The researchers use bagging in order to address multiple problems arising from high collinearity in logistic regression. Two problems arise from input variable collinearity. One of the problems is that high variability in the credit estimate makes the model difficult to interpret. The other problem is that high collinearity causes strong input variables to suppress weaker ones (Hastie et. al., 2009).

As mentioned earlier a problem with Shao and Li's model - by their own acknowledgement and Dalessandro's et. al. - is collinearity. This can cause the coefficients to be negative or overly weigh strong variables over weaker ones. In the context of mobile advertising this is can be seen as an increasing issue. This is because as the targeting algorithms of different channels become more sophisticated, they will target the limited amount of valuable users with a higher propensity. This in turn means that seeing ads from different channels will be correlated; on average channels will see the same users as valuable and target them. However, collinearity can be somewhat mitigated with bagging and misclassification rate and standard deviation metrics provide visibility to potential problems arising from collinearity. Another problem with the model is that, unlike the Shapley value model, it does not take into account the sequence of ads the user sees (Abishek et. al., 2012) – a search ad might be more effective when a user is already exposed to the product in earlier ads.

Finally, the researchers found that using bagging gives a similar misclassification rate as a normal logistic regression model but with a much lower variability. This to me, proves that using bagging with logistic regression is an improvement on a simple logistic regression model when it comes to MTA because this increases confidence in the model's consistency. The researchers also highlight selecting the right variable dimensions to decrease noise and make the outcome easier to understand. Having too much variable dimensions and elements in a set could require a lot more data or introduce so much noise to the outcome that it becomes very difficult to make decisions on. The domain knowledge of marketers should allow them to pick out the correct input variables for the MTA model.

2.3.5 Hidden Markov Model

Abishek et. al. (2012) developed a Hidden Markov Model (HMM) to attribute credit to different touch-points in a conversion funnel. In addition to attributing the credits of a conversion their model also allows for not attributing a conversion credit to ads. This means that it can also consider the effect of offline advertising activities. The model takes into account situations where increase in conversion probability caused by offline (un-observed) touch-points is not attributed to online touch-points. This is an improvement to the two previous models, which only divide the credit of conversions between online touch-points.

A Hidden Markov Model is a version of a Markov Model where the system that is modelled is a Markov Process with hidden states. A simpler Markov model is the Markov chain in which all states are visible. A Markov chain models a sequence of possible events of which the probability is determined only by the state of the previous event. A process is a Markov process if predicting the future of it can be done only by knowing its current state, i.e. knowing the states before the previous one does not affect the prediction. A simple example of a Markov chain would be predicting the amount of money you have after a coin toss. Knowing if you won or lost the previous tosses doesn't matter, the only thing that matters is knowing how much money you have before the next toss.

In the context of this thesis, the prediction of the coin toss would be whether the user installs the game, and the current state would be the level of the conversion funnel the customer is at. As it is impossible to actually know how far a user is in the conversion funnel or if the user is in the funnel at all - the user may never convert - a HMM is used. This way knowing the state a user is in is not needed. The model estimates the value of a touch-point as a function of its impact in the future but also the earlier touch-points before the current touch-point. The Hidden Markov model differs fundamentally to the logistic regression and Shapley value models mentioned previously. The Markov model reflects the different states of the conversion funnel a user is in and allows a user to move fluidly from one state to the next – even backwards in the funnel.

Another good attribute of the Markov model is that it takes into account the time dynamics of touch-points. When testing their model against Dalessandro's model, the LTA model and two other models Abishek et. al. find that the conversion root mean square error is the smallest using

their model. This means that their model had the best fit with the test data out of the five models compared. Unsurprisingly the worst was the LTA model. What is surprising is that compared to the other attribution models the Hidden Markov model attributed a lot less credit to advertising activity in general. This highlights another strength of the model, where other models always divide the credit between advertisers the Markov model also takes into account whether or not a user would be likely to convert anyway even without paid touch-points.

2.3.6 Bayesian MTA model

Another way of modeling MTA is by using a Bayesian framework (Li and Kannan; 2014). Li and Kannan study MTA on a dataset from the hospitality industry and approach the conversion funnel as a three-stage process. The first stage is a consideration stage where a user realizes their need and chooses a channel mix from which to gather information. In the second stage, the user visits the target website from a channel chosen in the first stage. In the third stage, the user makes a purchase. The researchers define a user's propensity to advance to the second stage of the funnel from a specific channel as the difference between the perceived attractiveness of purchasing through the channel and the cost of finding the required information. Their framework also sees spillover effects as coming from prior visits to the target website from the different channels. Li and Kannan model spillover and carryover on how they affect the cost of visiting a channel. They see a purchase coming from multiple visits to the target website, with each visit increasing the probability of conversion. To reflect this view Li and Kannan build a three-level MTA model.

Each level of the model reflects one stage of the conversion model in the researchers' framework. The first level of the model assumes that users are heterogeneous and allows for different consideration sets of both user and firm-initiated channels. In the first level of the model, the consideration set and the latent utility for each channel in the user's consideration set is defined. Next the channel visit decision is modelled by calculating both the cost and perceived utility of visiting a site, the lag effects of prior visits and the time since the previous visits. In the last level of the model the purchase decision is calculated assuming the general value of purchasing a product varies by individual along a certain mean, with the mean depending on the channel through which the users came to the target site.

I think that the framework Li and Kannan used to base their model on is very specific to their data and is not very applicable in the context of this thesis. Their framework puts a lot of focus on the number of visits to the target site and have a very different view on how a channel creates value to that of the other studies in this thesis which view the value added by a channel as an increase in conversion probability. The framework also assumes a conversion funnel where users do more research on the product through different channels. This is not the case in mobile gaming advertising where most channels and touch-points are not user initiated and where the research done before installing a free game is not very high. As such I think that Li and Kannan's model might fit a better for a product that people usually do a lot of research on and put some weight to where they search for information – e.g. buying a car. Finally, I did not find that the model satisfied the requirement of interpretability very well as the assumptions that the model's stages were not very easy to understand.

2.3.7 Aggregate level MTA

The previous models have used individual level data to divide conversion credit between touch-points. However, aggregate level data can also be used to build an MTA model to allocate budgets between channels. De Haan et. al. (2016) use a structural vector autoregressive model (SVAM) on aggregate level data. They study how relative effectiveness differs between online advertising platforms, how long the advertising effect lasts and where the effects have most impact in the funnel. Their study is close to that of Li and Kannan's but differs in the aggregate-level methodology and in that, they take into account offline advertising actions. The data that De Hann et. al. use is from an online-only retailer. The data is at daily product category level, with variables like daily cost of certain advertising type and daily number of site sessions started per category.

The model that the researchers use is designed for time-series data on the aggregate level. This model has four steps. In the first step, variables are tested for temporal Granger-cause, i.e. test whether or not one variable's time series has causation for another variable's time-series with e.g. t-test. The variables that have causation between them are classified as endogenous to the model and ones that are not are classified as exogenous. In the second step the form of the variables to be included to the model is determined with co-integration and unit-root tests. In the third phase, SVAM is used on the outcomes of the first two steps. In this phase dynamic models are estimated using the data from the first two steps and the Bayesian information

criterion. The vectoral autoregressive model built within this step is turned to a structural vector autoregressive model by applying restrictions to it that restrict the influence that changes in the lower funnel levels can have on the upper levels. Lastly restricted policy simulation is used to answer which channel has the highest revenue elasticity i.e. the channel that has to most impact on revenue. The last step allows estimating what effect an impulse has on a marketing action.

The researchers estimate that the company that provided the data for their research could improve its revenue by 21% by changing its current budget allocation to the one suggested by their model – which seems like a dubiously high number. The model's strengths are that it is very easily interpreted and that it uses aggregate level data which is easier to acquire and requires less computing power. One caveat of the research is that the channels - email, search, television, radio - it attributes credit to, have very different formats and are on different levels of the funnel. With mobile gaming most of the channels are very similar which could affect how well the model performs.

2.3.8 Mutually exciting point process model

One way to model the effects of touch-points beyond last-touch is to use a mutually exciting point process model that Xu et. al. (2014) develop in their research. By exciting point process, the researchers mean a process where touch-points stimulate the effects of the proceeding touch-points. The model takes into account the diversity of users, ad effect time decay and touch-point formats as a multivariate stochastic model incorporating user heterogeneity with a Bayesian hierarchical model framework. The mutually exiting stochastic nature of the model means that it assumes that the probability of different random points in time (touch-points and conversions) is affected by the points that happened earlier. Also, the effects of the occurred points are modelled to be decaying over time.

The data the researchers use for the model consists of impression and click data and conversion data. The clicks in the data are of website visits and search ad clicks and the researchers find that impressions without clicks rarely lead to conversions. The click-touch-points in their study are user initiated. This is very different to the context and the majority of touch-points in the data of my study, which are mostly video ads and served to the user without the user's direct input but leading directly to conversions. I think the assumption of the touch-points being

random points in time is valid but no the assumption that the earlier touch-points affect the probability latter ones happening – though the assumption might hold for search ads.

2.3.9 Search behavior and offline advertising

A challenge with MTA models is modelling touch-points that happen offline. Usually offline data is very hard or impossible to collect at individual level. This means that an MTA model that includes offline advertising needs to consider offline touch-points at the aggregate level as touch-points affecting all users. This is not impossible, for example the Hidden Markov model suggested by Abishek et. al. takes into account the effects of offline touch-points to some degree. A research paper by Joo et. al. (2016) touches on this subject. They research the effect that TV advertising has on keyword search. This is achieved by developing a three-level conditional choice model trained on search data of financial services and hourly spend on branded TV ads.

The three levels model three different behaviors that likely would be positively correlated with TV ad impressions. The first level models the share of users that search for a financial services keyword. The second level models how users choose either branded keywords or generic keywords. The last level models the click behavior of users. What the research finds using this model is that there is nearly no evidence that TV advertising is related to the number of users who search for financial products. This points towards that search propensity is caused by user level reasons. However, the model does indicate that brand searches are positively correlated to TV advertising, indicating that even though TV ads do not initiate searches, they do drive searchers to the advertised brands. However, the carryover over of TV advertising lasted only a couple of hours. The researchers estimated the effect that TV advertising had on search touch-points by using elasticity calculations where they calculated the change in the number of searches per brand when spend on TV advertising for each brand changed by 1%.

The model Joo et. al. propose estimates the effect of offline advertising has on online touch-points. Their model explains 70% of the variation in dependent variables. However, this model only estimates the effect offline touch-points have on online touch-points and does not estimate the effect online or offline touch-points have on other touch-points that are on the same level. In addition, the model only estimates the increase in number of searches and not the click-through rate on the search advertisement or video ads.

2.3.10 Empirical generalizations on MTA

Anderl et. al. (2016) look at what empirical generalizations can be made on MTA on the marketing industry level. They use a Markov model based on Markov chains on a large set of user level data from companies from travel, fashion and luggage retail industries. To validate their model, the researchers look at the variability and the predictive power of the model. Even though attribution is a metric that looks to the past, the model it's based on should also be able to predict conversion probability as it assesses how important different touch-points are in creating a conversion. To assess the accuracy of different attribution models they use ten-fold cross-validation and a receiver operating characteristic curve simplified to a scalar value. To gauge the stability of the attribution models Anderl et. al. look at their standard deviation of the predictive performance for different data sets and the standard deviation attribution results.

The generalized findings of the research is that touch-points initiated by the company are undervalued in simpler attribution models like LTA and user initiated touch-points are overvalued. However, the simpler models don't consistently overvalue the credit allocated to search engine optimization (SEO) and industry and brand affect the value added by SEO.

2.4 Literature review summary

The literature review has touched on how advertising affects people on the individual level, how the effects of advertising decays overtime and the different models used to attribute conversions to multiple touch-points. Research into how ads affect people has found that ads can affect people by repetition through which people learn and remember. Many different frameworks have been suggested to approach a user's purchase decision, from consideration sets to planned versus impulse buying. With mobile games the decision to download a free game is a low risk decision and as such I think it is more relevant to approach the install decision as an impulse decision. The ads affecting a user's decision state are touch-points in the mobile ad system which was explained briefly.

The main focus of the literature review was on previous research on multi-touch attribution models. Three dimensions have been proposed by Dalessandro et. al. to assess how good an

attribution model is: fairness, data-driven and interpretability. All of these are important if an attribution system is to be implemented among all stakeholders. To assess how well an attribution model reflects the effects different touch-points have on a user's probability to install, researchers use trained models to predict the conversions probabilities of users in a test set.

Multiple different models have been proposed to estimate MTA. The Shapley value is a collaborative game theory-based model that Dalessandro et. al. and Yadagiri et. al. use for MTA. As an MTA model Shapley value is easily interpretable, and it can take into account the organic probability of users installing without any touch-points. However, a major drawback of the Shapley value is that it is difficult or impossible to take into account ad time-decay and the effects of offline touch-points. One of the first MTA models proposed was a bagged logistic regression model by Shao and Li. This model does not take into account the order of the touch-points or offline touch-points. A hidden Markov model was proposed by Abishek et. al. who claim that it over performs the Shapley value models. The hidden Markov model takes into account time-decay and offline touch-points.

Multiple other MTA models have been proposed. Many of the models have assumptions that are based on frameworks that I do not think are applicable in the context of this thesis. What the models have in common is that all beat the LTA model and find that the LTA model over-estimates the contribution of touch-points lower in the funnel.

3 Data

This thesis uses user level impression and install data for both converted and unconverted users for a single game during a 37-day period. The data consists of 44,719,217 users that have seen marketing messages through eight different channels on two platforms (iOS and Android). Users who have not seen the company's ads within the time period are not observed – thus so-called organic users and their install probability can't be studied.

It is worth mentioning that it is almost certain that some observed users have seen marketing messages before this period. This is a challenge as it means that some users have recently interacted with touch-points that are not recorded in the data but that still affect their conversion probability. This in turn will probably skew the data so that the user cohorts observed at the beginning of the observation period will have on average a higher probability to convert than the later cohorts. In addition to this, the last cohorts may have lower conversion rates with similar amount of touch-points than the earlier cohorts due to having less time to convert. Another caveat is that the impression data from three channels is not available as they do not share impression or click data at the user level. This means that there is no visibility to the touch-points observed users have had with these channels.

The data used by this thesis is made from combining two different datasets. One dataset has user level impression data and one dataset has device level install and spend data. The impression data is organized so that one row represents an individual impression for an individual user. Each row has: the channel through which a user saw the ad, the game that was advertised and a hashed ID through which the impression can be matched to an install or other impressions of the user. Each row also has the impression time as an ISO 8601 time stamp, meaning the impression time is reported down to the second. However, only the date level is used in this research. This is because the install data is at the date level and because the impression data is from a different source, so the time-zone doesn't match with the install data.

Table 1: impression and install data columns

Install data columns	Install time	Hashed user ID	Money spent	Platform	Game
Impression data columns	Impression time	Hashed Device ID	Channel	Platform	Game

The install data identifies users at the device level and not at the user level. This means that if a user saw an ad and installed the game and then installed it on another device the user will have two conversions. This is not optimal as it can be argued that the user has already converted when installing the game on a second device. On the other hand, if the user level id would be used there would be a lot of users who have an install time dating back years, which would make using the data difficult. I also assume that not many observed users installed the game on multiple devices during the data collection period, so I assume the effect of these users on the outcome on the model is small. The installed devices are matched to the user impressions through a third table with both the hashed device IDs and hashed user IDs. When matching the installed devices to the users some users might have multiple devices registered. To minimize the effect of this, only device IDs that have the install date after the collection period are counted. If a user still has multiple different device IDs the device ID with the latest install date is chosen. Only App-opens are counted as installs and the first app open date is marked as the install date.

As mentioned earlier there are some problems with using data from the beginning and end of the data collection period. To address the problem, I will omit the first seven days from the impression and install data and the last seven days from the impression data. This means that if a user installs on day three of the gathering period, his installs and impressions are not used in this thesis. The installs of the last seven days are counted only if the users did not see an ad during the last seven days. In order to have the paths more comparable, the touch-points per user are limited to ones that happened within seven days from the last impression or an install. The seven-day time period is used because this time period is assumed as the time the ad effect time decays to zero. More on this in the next chapter.

Finally, the users are divided into two cohorts: training cohort and test cohort. The randomly drawn training cohort will consist of 70 % of all applicable users - i.e. users who are observed within an appropriate time frame within the data gathering period. The rest of the applicable users are in the test cohort.

Table 2: general information on user cohorts

Installs from paths with one impression	Installs from paths with more than one impression	**Total Installs**	Installs from paths that need MTA
194968	194568	**389536**	37184
50.1%	49.9%	**100%**	9.5%

From Table 2 we can see that the amount of installs that come from paths that have impressions from more than one channel during the attribution window is only about 10% of the total installs. This is very important. If the amount of installs that would need an MTA model is only one tenth of the total number of installs the benefit derived from an MTA model could prove small.

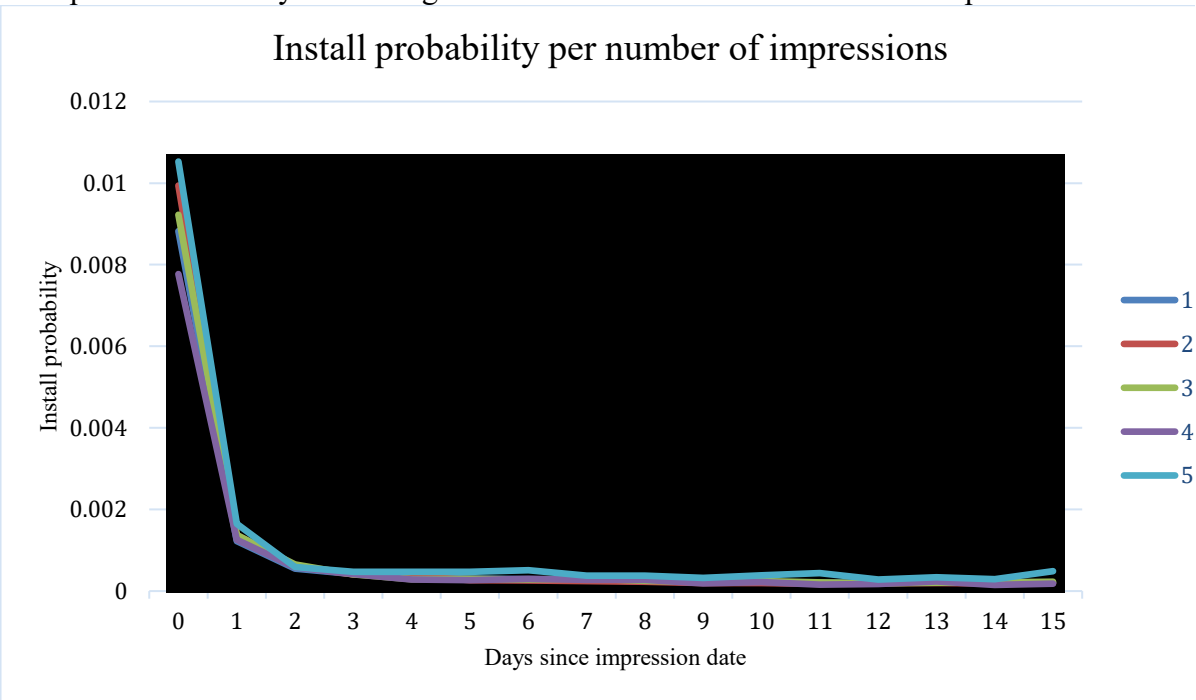
3.1 Ad effect time decay

To determine the period within conversions should be attributed to touch-points, I look at how the conversion probability develops after users saw ads on a single day. This was done by taking the users who only saw ads on a single day. The impression data was then aggregated by user and the impressions counted so that there was only one row per user with the number of impressions the user saw and the date of the impressions. Then these users were grouped by their impression numbers so that for each day there was the number of impressions per user and the number of users who saw that number of impressions. This was also done with the install data so that for each day's impression cohort there was the number of installs per impression number per day since the cohort's impression date. Each day's aggregated data was then aggregated further together so that the number of impression users was summed per number of impressions seen per user and the install data was summed per number of impressions seen per user per day since the impression happened. This aggregated data was then used to calculate the probability of conversion per day since the impression date per impression class.

Table 3: Install probability by days since impression date and number of impressions seen per user

Days since impression	Amount of impressions per user 1	2	3	4	5
0	0.008822	0.00994	0.009224	0.00777	0.010528
1	0.001229	0.001373	0.001372	0.001259	0.001649
2	0.000553	0.000603	0.000655	0.000585	0.000587
3	0.000414	0.000407	0.000408	0.000414	0.000474
4	0.000338	0.000373	0.000294	0.000286	0.000472
5	0.000311	0.000268	0.000298	0.000265	0.00047
6	0.00029	0.000264	0.000283	0.000299	0.000511
7	0.000273	0.000244	0.000306	0.00028	0.000379
8	0.000234	0.000231	0.00025	0.000276	0.000379
9	0.000222	0.000201	0.000218	0.000194	0.000326
10	0.000204	0.000199	0.000244	0.00022	0.000385
11	0.000201	0.000179	0.00021	0.000155	0.000438
12	0.000202	0.000194	0.000191	0.000174	0.000284
13	0.000215	0.000201	0.000207	0.000235	0.000339
14	0.000215	0.000187	0.000212	0.000152	0.000295
15	0.000206	0.000197	0.000226	0.000182	0.000484

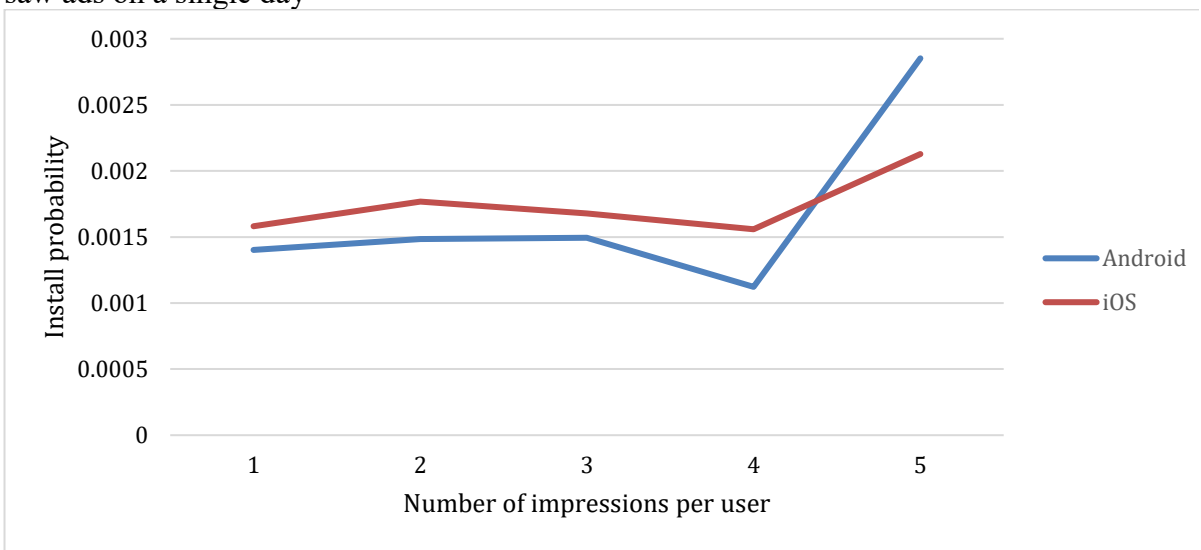
Figure 1: How install probability develops per amount of impressions seen per user on a single day as time goes by from the impression date. Each line represents a certain amount of impressions seen by users – e.g. the red line is for users who saw two impressions.



As we can see from Figure 1, the install probability is on average almost zero (~0.0002) on the seventh day after seeing an impression for people who saw 1 to 5 impressions. For the cohort of people who saw five ads in a single day the conversion probability fluctuates more but this is probably due to statistical noise due to the small cohort size. When dismissing day zero from the graph, it is easier to see that the conversion probability decreases further after day seven. However, the install probability on day seven is so small that I would rather use the industry standard seven-day attribution window and have bigger cohort sizes to look at than block out even more days at the start and end of my data collecting period.

In Figure 2 I've taken users who saw ads on a single day and looked at their conversion probability within the following 7 days. From it, it appears that there is a positive correlation between number of impressions a user sees and install probability. However, Figure 2 holds a bias. Users who play a lot might have a higher probability to install just because they consume more gaming content on average. Because users who play a lot run into more ad placements, they would also have more impressions thus biasing higher install probability towards higher number of impressions. As such I'm not very confident to state that a higher number of impressions correlates to higher install probability based on Figure 2.

Figure 2: How install probability develops as impression amount increases for users who only saw ads on a single day



To look at the conversion probabilities of users who saw ads on multiple different days, I assumed the ad effect to decay to zero after seven days. With this assumption I looked at the impressions that happened in the last seven days before the last impression a user saw or before the user installed. I excluded the users whose last ad happened in the first seven days and last seven days of data collection period. I also did this for installs. The reasoning behind this was that if an ad affects a user for seven days, a user whose last ad or install was before the seventh day of the observed period might have seen ads that affect their install probability before the observed period. This same reasoning applies to the cohorts at the end of the observation period. Users who saw ads e.g. three days before the end of the period would have lower probability to install because they would have less time to convert. The impressions that happened over seven days before the last impression or install were also excluded as their ad effect would be zero. Because I look at the last seven days before last impression or install, this data also includes some of the users in the earlier discussed table and Figures. Figure 3 depicts how the install probability develops as the number of impressions within the seven-day period increases. In order to reduce noise, I separated the networks and platforms so that I'm looking separately at each network per platform.

Figure 3: How Install probability develops per channel on iOS as the amount of impressions a user saw in 7 days increases

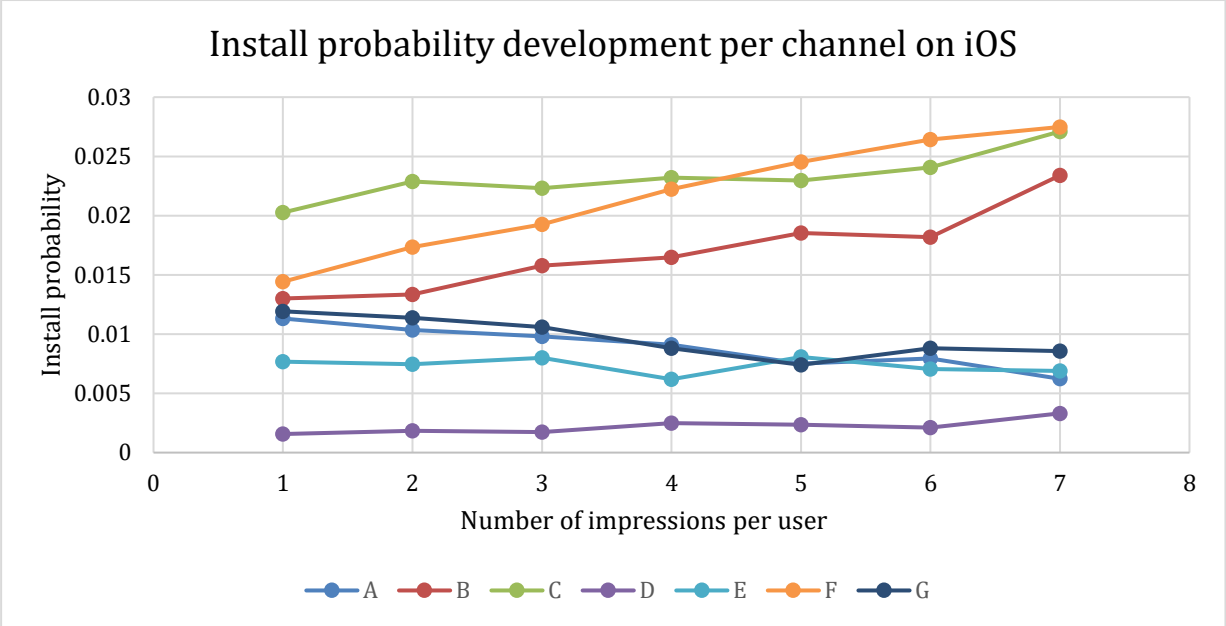
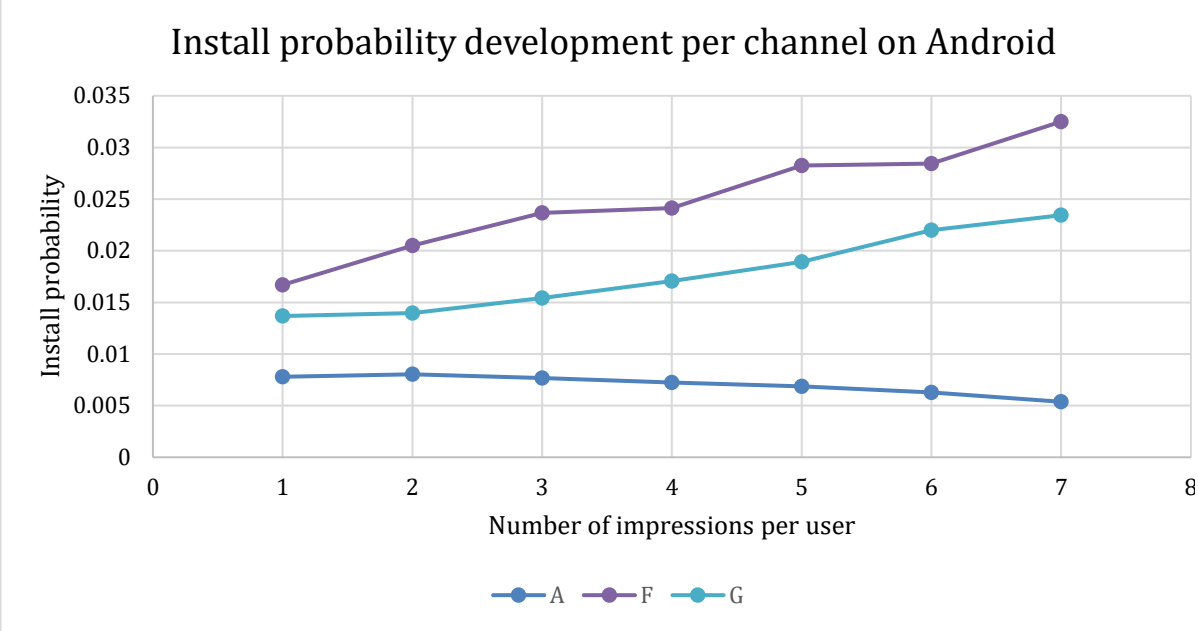
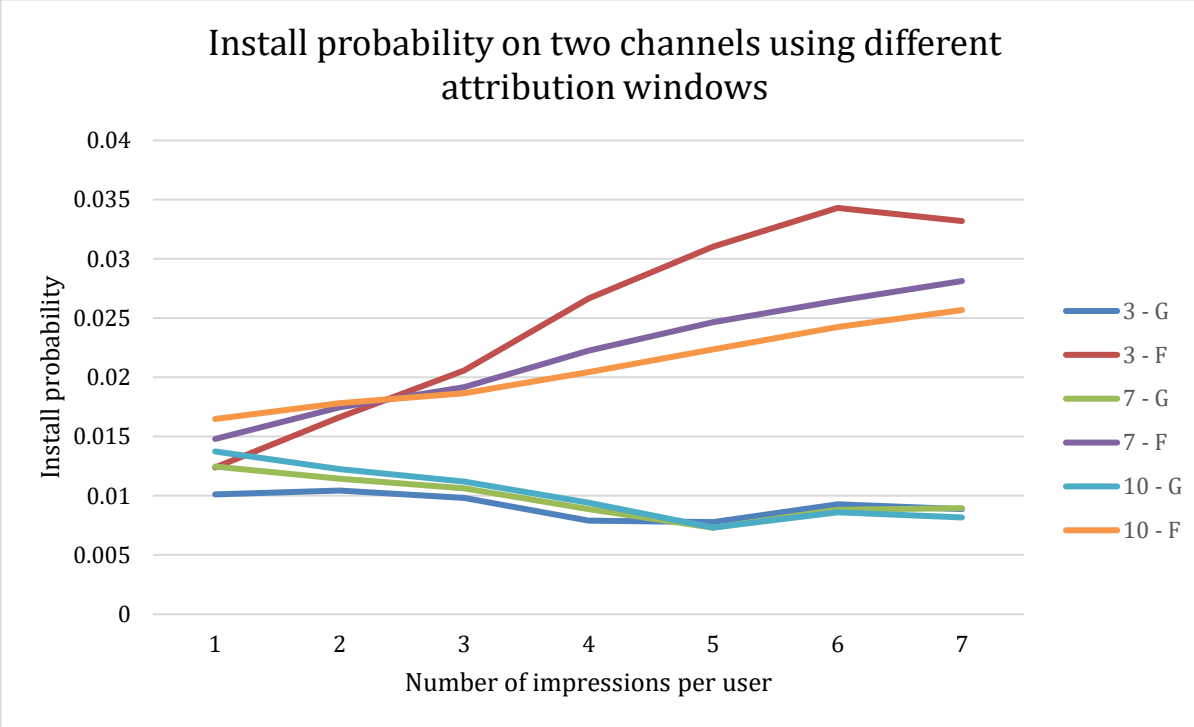


Figure 4: How Install probability develops per channel on Android as the amount of impressions a user saw in 7 days increases



From Figures 3 and 4 we can see that how install probability develops differs a lot from channel to channel and in the case of channel G, between platforms. This is very interesting and there are probably many reasons why the Figures look like they do. One reason might again be biases. Because I'm looking at a seven-day window people who don't play that much might not make it to the higher impression amounts within the seven-day window and if they do not install, they are dropped to the lower impression cohorts skewing their probability. Another reason for the differences might be due to differences between channels. One channel might bombard users randomly with low quality placements and another might use more precise targeting. This highlights one of the reasons an MTA model could be an improvement on the LTA model, different channels likely cause different uplift due to their different targeting methods and placements. Finally, to gauge the effect of the time window I looked at how the install probability for different channels develop as the time window is changed. Figure 5 depicts two networks on iOS. It looks like the install probability develops the same way even with different time windows. This makes me believe that the time window does not distort my findings too much.

Figure 5: How install probability changes as the time window changes. Legend letter depicts a channel and legend number depicts attribution window used. 3-G means channel G install probabilities when using a three-day attribution window.



We have now looked at the ad effect time decay and at how the install probability develops as impressions increase. When looking at the ad effect time decay, we can clearly see that the effect of an ad decreases rapidly as time goes by. This supports the assumption I made earlier that the purchase decision would be an impulse decision. If it were a more planned decision, I would expect the probability to decrease much more slowly. The correlation between the number of impressions per user and install probability seems to at least partly depend on the channel. There is no tipping point after which the install probability starts to trend downwards for the channels where the probability seems to be positively correlated with the number of impressions.

4 Methods

Different models can be used to estimate the credit channels deserve installs. Some of these models have been discussed in the literature review. To divide credit to different channels this study will apply the Shapley value model and similar model that considers time and ad effect time decay. These models were chosen because they satisfy the three measures defined by Dalessandro et. al. (2012). They are easy to interpret, they are data driven and they are supposedly fair. It is difficult to say whether a model fits in with the last measure because I interpret fairness as being the same as how accurate the model is in estimating the uplift a touch-point generates. This can't be directly observed and will need to be tested – more on this later.

In earlier research Abishek et. al. (2012) found that a Hidden Markov model was more accurate than a Shapley value model. However, I do not see that it fits well in the setting of my data. The Hidden Markov model used in the study estimates the probability that a user moves through different states before converting - the states being different parts of the funnel. This might make sense in a different setting where the touch-points are in different levels of the funnel and one touch-point might send a user to a touch-point lower down the funnel. However, because user acquisition in mobile games is direct response marketing the purchase decision is quick and low risk. Because of this touch-points are on the same level of the funnel and a Hidden Markov model does not fit best. In my data users move from one touch-point to another at random independent of a user's state. A binary HMM where one state is un-converted, and one is converted could make sense but as the Shapley value model has already been used in a MTA research it is an easier place to start from with my data.

The models are trained with a 70% training set after which the model will be used on a 30% test cohort to predict different user journeys' install probabilities. The same is done with the LTA model to see how they compare. This 70-30 split is some-what arbitrary but it should ensure that the models are acceptably accurate, and that the test data touch-point paths have large enough cohort sizes.

4.1 Shapley value model

As mentioned earlier in the literature review, Shapley value is a value generated by a collaborative characteristic function game and is used to divide the credit of the outcome of a game to its participants. Shapley Value assumes a game where there are multiple different players all cooperating towards the same goal. Shapley value helps allocate a fair share of the credit to the participants by the contributions the participants make towards the common goal. In the context of this study the players are different channels, the outcome is the install probability and a game is the aggregate of the paths that have the same amount of touch-points from different channels.

Shapley Value function calculates the credit a player deserves from a collaborative game for all permutations of the game the player can be in. Then the function calculates the average credit of the player from all the permutations the player is in. This is done for all players in the coalition. To further clarify: a when talking about a player I am referring to a player as a participant in game theory's collaborative game, I am not referring to a player that plays a mobile game – I refer to them as "users". In the context of this study, the function would count the following:

1. Calculate the uplift in install probability – i.e. the value - a touch-point deserves and count the credit a channel deserves from its touch-points in a single path
2. Calculate the average credit of each channel across all permutations of the path

To calculate the install probability of a path, users need to be grouped by their paths and the number of installed users per path needs to be divided the by all users per the same path. With the install probabilities of different paths, the uplift a touch-point adds to the path can be calculated. This can be done by looking at the sub-paths that make up the whole path. If the touch-point we want to calculate the uplift for is X_i , then the path up until i is X_{i-1}. Because the cohort size is large for the training set, most paths are observed, and their install probability is known. To calculate the contributions of different touch-points I start from the touch-point closest to the install and work my way to the oldest touch-point.

Because the install probability of a path where i is the last touch-point and the install probability of the path X_{i-1} is known it's easy to calculate the uplift i provides:

$$U_i = P(X_i) - P(X_{i-1}), \qquad (2)$$

where U_i is the uplift provided by touch-point i.

In some cases, the $P(X_i)$ is smaller than $P(X_i)$, in these cases the uplift a touch-point provides would be negative. This does not make sense because that would entail that the ad a user saw made them less likely to install. This might be true in some cases but in the bigger picture I find it more likely that this is due to statistical noise or selection bias – users who have not installed after seeing many adds are unlikely to install after additional ones. Therefore, if $P(X_i) - P(X_{i-1})$ is less than zero I adjust the uplift to be zero. Another edge case is when a sub-path can't be found, in such cases the uplift for some of the touch-points cannot be calculated. In these cases, I credit zero uplift to the touch-point that uplift cannot be calculated for and set the probability of $P(X_i)$ to be the same as the last known sub-path. Luckily these unknown sub-paths only happen when the parent paths get very long and have very few users who go through them. With very few users the effect of this edge case is probably small.

The contributions of touch-points in a path are grouped by the channel they belong to and summed. The sum is the channel's contribution to the path. Once channels' contributions are calculated for each path the paths are grouped together by the channels involved, the number of the channels' touch-points and in when taking time into account the impression times. For these groups the average contribution of the involved channels are calculated and their share of the total contribution of the group calculated. This share can then be used to attribute conversions when a converted user's path consists of touch-points from multiple different channels.

$$C_{ap} = \frac{\frac{\sum_{p=n}^{n} U_{cp}}{n}}{T_{ap}}, \qquad (3)$$

Where C_{ap} is channel C's share of credit for all paths in path group ap, U_{ci} is the uplift channel C provided for path p, n is the number of paths in path group ap and T_{ap} is the sum of average contributions of channels in path group ap.

Briefly explained, Function 3 first sums the uplift channel C provided to all the paths in path group ap, it the divides the sum by the number of paths in the group to get the average contribution of channel C. The average contribution is then divided by the sum of all channels' average contributions in the path group.

The benefit of this model is that it takes better into account the number of touch-points a channel has in a path. So that if the ad effect time decay was slow and the effect of ads from different channels were equal the channel with the most touch-points would get the most credit. It also considers the quality of ads better than the LTA model. If there was one channel that spammed small banner ads everywhere, an LTA model would attribute a lot of conversions to it just because every now and then the channel would manage to slip an ad between an impactful ad and a conversion. Because the Shapley value takes into account the uplift a touch-point generates, ads like these would not get that much credit because a path with a banner ad would have a conversion probability very close to an identical path without a banner ad.

One problem with the normal Shapley value is that it does not take into account timeseries data. If time series data is not taken to account, an impression that happened on the day of the install would have the same install probability as one that happened seven days ago – from what I found earlier, this clearly is not the case. A solution for this is to break the touch-points down to the daily level so that one player is a channel's touch-point x days before the install. The problem with this is approach is small cohort sizes. The more granular the players get the smaller cohort sizes get for paths.

The credit attributed to each touch-point represents the value a channel added in the conversion funnel. This value can be thought as an increase in the probability of a user to convert. In other words, the credit attributed to a channel and its install probability uplift can be thought as two sides of the same coin. As attribution credit and probability are tied together, we should be able predict - to a certain degree - a user's conversion probability based on the touch-points they have interacted with.

4.2 Time decay model

Where in the Shapley value model channel contributions were averaged by grouping similar paths, the time-decay model looks at channel contributions at the path level. The model breaks down the paths by the time an impression happened. This means that a single touch-point path is broken down to 7 different paths based on the how many days before the last impression or install the touch-point happened. The amount of days from the touch-point until the last impression or install are considered to calculate the share of credit a touch-point deserves. A touch-point's probability is calculated by filtering the ad effect time decay data for install probability by the touch-point's channel, platform (iOS or Android) and impression time. Once this is done for all touch-points the share of credit is calculated by dividing each touch-point's install probability by the sum of all touch-point probabilities in the path. This means that if a path has three touch-points from a same channel on the same day, they are each allocated one third of the credit. After the share of credit per touch-point is calculated they can be grouped by channel and summed to get the channel's share of credit.

The problem with this method is that it does not consider the path's install probability. It only takes into account the ad-effect time-decay probabilities. This model also does not take into account the possible dependencies between touch-points. For example: if two channels have synergy together this model could not divide the credit of this added benefit between the touch-points. This being said, I believe that synergies have a very small weight in the setting of this study. This is because the marketing campaigns are direct response campaigns where the channels are on the same level of the conversion funnel. With bigger purchases such as cars the funnel is usually very deep with a lot of different channels - search, website, dealerships - on different levels of the funnel, dependencies would probably have a significant weight. With the touch-points on the same funnel level, the model should be quite good at spotting differences between channels' affects. If a path had two touch-points on the same day – one a video ad and one a banner ad – the model would take into account, the install probabilities on day zero for both channels and the banner ad would get less credit. This model is also very easy interpret and it's easy to understand how the results were generated which satisfies the third property of a good MTA model.

5 Model comparison

Validating an attribution model is very difficult. The only real way to test whether one attribution model is better than another is to make an A/B test based on ROAS. In such a test an audience is divided randomly into two groups and only shown ads from the campaigns in their assigned groups. Each of these groups would have a different attribution model to attribute credit and help a marketer allocate budget. The attribution model in the group producing best ROAS would be the winner. Alas, this setup is technically not supported by most of the channels in the mobile advertising ecosystem and falls outside the scope of this thesis. Because there is no ground truth to observe, any other way of validation is directional and used to weed the attribution models that are most likely to perform badly.

To validate the proposed models, they are compared to the LTA model on how well they can predict the install probabilities of paths in the test data. A good attribution model needs to reflect the uplift in conversion probability caused by a touch-point as accurately as possible because an attribution model is used to give credit to touch-points based on their effect on a desired outcome. A perfect attribution model would thus know exactly how much of an install was caused by the touch-points and how-much of it was organic install propensity, it would know the exact uplift different touch-points provided. With a perfect attribution model, we could thus predict the exact conversion probability of user-paths.

I will use this property of attribution models to validate my MTA models against the LTA model. I will look at the test sets user-paths with time. This means each touch-point will have information on the amount of days between it and the last impression/install. Zero days means the impression happened on the same day as the last install or impression. I will look at how far the probability predictions generated from the attribution models are from the actual install probabilities of the test set. To decrease noise from the validation I will only look at paths that have at least 1000 total users. This will decrease the number of paths the validation covers significantly. However, the remaining paths hold most of the users and there should still be enough paths to get a statistically significant result.

Each of the four variations of attribution models – including the LTA model- will have its own setup on how to generate a probability prediction. The timeless Shapley value model will compare the average install probability of the aggregated path groups to the paths in the test set

that would belong to the aggregated path group. However, the test cohort's paths will also include time data. The Shapley value model with time data will be compared to the test data much in the same way as the model without data except its grouping also considers time series data. This approach is justified because the Shapley value model calculates the contributions of channels by calculating the average of their contributions to a group's paths. Thus, the probability generated by the model should also be the group's install probability.

The third model is based on the ad-effect time decay data. At first glance it would seem logical to test this model by comparing the training set's time decay data, as this would show how well the model can predict an ad's effect at day x before an install. However, this approach would not be a fair comparison to the other models that need to predict the probabilities of each individual path and would probably always lead to the time decay model being the winner. Producing an install prediction for this model is a bit more difficult than to the Shapley value models. The time decay model always attributes credit to all touch-points in path if their channel's ad effect has not decayed to zero within the time period. This means that in practice the model can't attribute zero credit to a touch-point even if it has not provided uplift to the path that it's in. It also doesn't capture any other uplift a touch-point might produce outside of its ad effect time decay. For example if a channel is not very effective on its own but very effective when combined with another touch-point; the model would not attribute it the credit it deserves because the credit would be based on its time decay data which is based on how well the channel works on its own.

Because the decay model assumes a touch-point's contribution based on its channel's ad effect time decay, the install probability prediction should also be based on the this, with all touch-points contributing to this probability. To generate probabilities for this model, the install probabilities of each touch-point will be taken from the channel's ad effect time decay filtered by how many days separate the touch-point and the install. In the actual attribution model the share of credit depends on how big the share of the touch-point's probability is from the path's total probability. To generate a probability prediction, this method will be used backwards. The total install probability of a path can thus be calculated by taking the probability of each touch-point, multiplying each probability by the share of credit the model has calculated for the touch-point and adding these multiplied probabilities together as the path's probability.

$$P(path) = \sum_{i=1}^{n} P(t_i) * S_i \,, \qquad (4)$$

where $P(path)$ is the generated path probability, n is the number of touch-points in the path, $P(t_i)$ is the probability of touch-point t_i and S_i is the share of credit the model attributed to the touch-point from the training data set

The LTA model is based on attributing all the credit to the last touch-point. This means that it assumes that all the uplift in install probability is due to the last touch-point. If this model were perfect, it should thus be able to predict the install probability of each path by the last touch-point. To validate this model, I will take the test path's last touch-point's channel and look at the what the channel's install probability is in the training data for a path with one touch-point taking time into account.

6 Results

In the beginning of this thesis four research questions were presented:

1. How well would a multi-touch attribution model attribute conversion credit compared to the last-touch attribution model?
2. What share of installed users have a touch-point from more than one channel?
3. How does install probability develop as impressions per user increase?
4. How does the effectiveness of ads decay overtime?

6.1 MTA installs, impression numbers and ad effect time decay

From the data it is clear, that only small minority of installs would need MTA to divide credit. Only 9.5% of all applicable installs during the data collection period has an impression from more than on channel during seven days before an install. There was also some evidence that the number of impressions per user was positively correlated with install probability for some channels. However, this is inconclusive. It is difficult to say decisively how the number of ads seen affect a user's install probability because the number of ads shown to people could not be controlled for. An increase in install probability as the number of ads increases could probably be partly explained by the fact that users who play more are more likely try many games and see many ads. Because the users were not divided into different groups in which the number of ads shown could be controlled, no conclusion can be drawn. This thesis also looked at the ad effect time decay i.e. how the install probability develops with time after users see an ad. The finding was that it decayed rapidly so that the effect was a fraction from what it was on day zero after two days. At eight days after impression the install probability almost zero and the attribution window was set to eight days.

6.2 Results from model validation

The results from the attribution model validation are somewhat surprising. The attribution model that could best predict the install probability of different paths in the test data was the Shapley value model that considered the time of impressions. The second-best model was the LTA model with the worst model being the time decay model. The key performance indicator for the validation model is the weighted average of the relative absolute difference between a

model's prediction of the install probability and the real install probability of each of the test cohort's path.

I chose this indictor because I see that it depicts fairly well how the models conform to two of the three traits of a good attribution model that Dalessandro et. al. came up with. By default, all the MTA models in this study are data driven so there is no use evaluating that. The relative absolute difference between predicted probability and real probability assesses how fair the model is from the real probability. The better a model captures the uplift of different touch-points the more accurately it can attribute credit to the channels deserving it the most, and the fairer the model is. If a model accurately captures the increase in probabilities provided by different touch-points it should be able to predict the install probability of a path by looking at its touch-points. Because of these aspects, the fairest model should have the least difference between its prediction and the true install probability of user paths.

Because an attribution model's purpose is to attribute installs, purely looking at paths can give a wrong estimate of the best model. This is because paths have different amounts of installs that they've generated. If an attribution model is good at predicting a zero-install probability for paths with none-to-little installs but bad at predicting the probability of paths with a lot of installs it is not as good as an opposite model. Put differently, when in real use, an attribution model is only used on installed users. To account for this, I use install weighted average to assess the relative absolute difference between predicted and true install probability.

Table 4: Weighted average of relative absolute difference between predicted and true install probability of all paths in test data

Weighted average of relative absolute difference between model's predicted install probability and true probability			
LTA model	Shapley value without time	Shapley value with time	Time decay model
12.4%	20.7%	4.2%	29.4%

Table 5: Weighted average of difference between predicted and true install probability of all paths in test data

Weighted average of difference between model's predicted install probability and true probability			
LTA Model	Shapley value without time	Shapley value with time	Time decay model
-0.045%	0.204%	0.004%	-0.418%

From Tables 4 and 5 it is clear that the Shapley value model that takes time into account can best predict install probability of the test set's paths and thus probably best captures the uplift provided by different touch-points in the paths. What's interesting is that the LTA model is better able to predict probabilities than the time decay model. I would have thought that the time decay dimension of the touch-points would have had such a large weight in predicting install probability that the time decay model would have fared better than the models that do not take time into account. Looking at the weighted average of differences in Table 3 it is easy to see that the time decay model under values the contributions of touch-points and the LTA model under values the contributions.

Part of this fits well with the discussion of the models earlier in this thesis. As I stated in the introduction of this thesis, the LTA model assumes that only the last touch-point has an effect in a user's purchase decision. This leads the model to under value the uplift provided by the touch-points that came before the last touch-point there by predicting a lower probability of install for paths with multiple touch-points. That the time decay model also undervalues the touch-point contributions is surprising. I would have predicted that it would over value the contributions as it cannot give zero or negative contributions to the touch-points. What might explain this is that the touch-points somehow complement each other. The time decay model would miss this effect as it does not observe the dependencies between touch-points at all.

One thing that affects the Shapley value model and that the validation method used does not capture, is the dependencies between paths in the model. As shown in Function 2, each touch-point's uplift is calculated as the difference between the install probability of a path and the install probability of the path leading up to the touch-point. Because the contributions for touch-points are calculated as the differences between install probabilities of the sub-paths any inaccuracies within the sub-paths will cumulate in the path. For example: even if the Shapley value model generates a very accurate install probability for a path, it will probably attribute credit for that path in-accurately if it generates very inaccurate install probabilities for the sub-paths of which the path is made of. On the other hand, if most of the probabilities for the paths that are generated by the model are accurate the model will probably be able to accurately attribute credit. In this light the Shapley value model that does not take into account times of impressions is probably even more in-accurate than it appears based on its relative absolute difference.

7 Conclusion

7.1 Summary

This thesis looked at how different attribution models, trying to find a model better than the LTA model. Multiple different models have been developed in earlier research. Out of these one was selected to be tested in the mobile gaming context. Three different models were trained and tested on large user cohorts of both installed and non-installed users. Two of the models are variations using Shapley value to attribute credit to different channels in a user path. One of the variations used time series data and the other one did not. The third model used the ad effect time decay of different channels to attribute credit. The models were compared to the LTA model by seeing how well they could predict the install probability of user paths in the test data set. Out of the four models tested the two best models were the Shapley value model that considered impression times and the LTA model. Out of these two the Shapley value based model was far superior, with a weighted average of the relative absolute difference between true and predicted install probability being just 4.2% compared to the 12.4% the LTA model generated.

In addition, this thesis looked at the ad effect time decay and found that the effects of ads decayed rapidly after the first day and was almost zero at day seven. The study also looked at how install probability developed as the amount of impressions increased. However, nothing conclusive was found. The most immediate actions to take from the findings of this study from a managerial point of view would be to test how different ad frequencies to users would affect the impression to install rate and overall ROAS. One way to test different attribution models would be to test different models between different countries but this kind of test would require time and would not be a perfect A/B test

7.2 Managerial implications

There are couple ways forward to test the findings of this study. The lowest hanging fruit is testing ad frequency. Earlier we saw that the amount, of ads users see correlates with their install probability. It would be useful to test different impression frequencies on different channels to see which one provides the best funnel conversion and thus lowest cost per install and the highest ROAS. This test would also be relatively easy to setup with audience suppression lists by adding users to a suppression list for a limited amount of time after they have seen an ad. This test would need to be based on comparing two different countries as channels do not have A/B testing capability.

The best way forward with MTA models would probably be testing them between two similar countries – e.g. Sweden and Denmark. Testing the attribution models would probably need to be done by applying it to the advertisers own reporting tools. This is because the mobile ad ecosystem for mobile games uses the LTA model and the channels are not setup to use an MTA model. In the current setup the bids per users are given per channel – the channel then uses this bid and the install probability of users installing to calculate how the advertiser's ads are shown. Using the model on internal reporting would mean that it would affect the allocation of marketing budget to different channels, but not the actual amounts paid for installs. Put differently, this means that the MTA model would allocate the credit of revenues of installed users to channels by their share of credit and not actually pay the channels based on attributed share of credit.

The total ROAS of all channels under each model needs to be compared to assess how well an attribution model works. Assessing an MTA model's effect on ROAS is difficult without a way to do real A/B tests. Testing models in different time periods is not possible because external factors like competition and changes to the game have a too high impact. Testing models between different countries that are similar is possible, but it can be difficult to find countries where performance is very similar, and which can support big enough scale long enough. Because testing an MTA model is mainly testing whether the model helps improve budget allocation the test needs to last for some time in order to allow the campaign manager to do adjustments to the budget allocation between channels. This combined with the fact that for most large gaming companies, user acquisition is a large investment means that the test would

probably need to be run for a long time for management to confident enough start using it. This in turn increases the potential cost of the test as it needs to be run at a large enough scale even if the test cells do not perform in terms of ROAS.

7.3 Limitations

There are multiple limitations with both the data and methods of this thesis stemming from how the data was collected and the methods used. First, even though the cohort size of this study is very large, it is observational and thus could not be controlled for many variables. Another problem with the data is that it is missing impressions from Facebook and Google which together got almost half of the marketing spend during the observation period. This means that a big share of the users observed, most likely also saw ads from these unobserved channels which influenced the users' install probability. Also, there was no observed hold-out group which would have not seen ads. Without such a holdout group it is hard to say what real uplift the touch-points provided to the organic install probability.

Another limitation of this study is the way the paths of non-installed users were generated. With the installed users it was easy to generate a path, take the install date of a user and search the seven days preceding this date for impressions. For the non-installed users this is trickier. For non-installed users I took the date of the last impression and looked at seven preceding days and matched these paths to similar paths in the installed users' cohort. The problem with this approach is that it distorts the paths' install probabilities. For example; if a user saw a single impression five days before installing, there would be no non-installed user path that would match to this. This is because non-installed user paths with only one touch-point the only impression happens on the day of the last impression which is day zero. This inflates the install probabilities of paths with a single impression that is not on the install date to 100% and drops the non-installed users to other paths - decreasing their probability.

A better way to combine non-installed users to installed users would be to look at each individual day and look at the paths of installed users in the seven days leading to the install day and look at how many non-installed users took the same path up to that date. Once this is done for all dates the same paths between different dates would be grouped together. However, there is a problem with this approach as well. If a non-installed user saw an ad on each day of

the data collection period, they would be applicable to be allocated to multiple different paths – should they be allowed to do this, or should each user be allocated to a path only once? The reason I did not implement this approach was due to the higher complexity it requires and time constraints.

This study focused on attributing installs to different channels but, this might not be the best event to measure. In performance marketing the key performance indicator is usually ROAS. In most high grossing freemium games, most of the revenue is generated by in app-purchases. As this study only focused on attributing installs, it ignores purchasers, who are more important than installs when it comes to ROAS. A fair MTA model attributing installs would offer most credit to the channel bringing in the most installs even if another channel brought more payers who generated more revenue.

I validated my models by using them to predict the install probabilities of the test set's paths. One problem with this is that I left out paths with cohort size less than 1000. I did this to decrease statistical noise brought on by the small cohort size. However, this means that the validation did not cover all paths. Another problem with the validation is the methods used to generate the install probability predictions from the attribution models. Any inaccuracies, biases and errors in logic in generating install probabilities from the models will have affected the validation itself. This means that even if a model were to attribute correctly it might appear in-accurate if the method generating probabilities from model was bad. Also, the validation model used in this study does not really show which attribution model works best. The bottom line is that the best attribution model maximizes ROAS. My validation model does not show us this. However, currently the mobile games marketing ecosystem does not have the products available to create a cross channel A/B test to test the ROAS different MTA models provide.

7.4 Suggestions for further research

There are several avenues for future research. One interesting topic would be trying to find the optimal amount of impressions or the optimal frequency to show impressions to users in order to maximize install probability. This would probably require controlling for the number of ads users see and would probably be best done with audience suppression lists.

There are a couple of interesting topics of future research when it comes to MTA models themselves. One would be to implement an MTA in a case study and measure the effects of different MTA models on ROAS. Another possible study would be to test additional models for MTA. One novel model to test would be a network diffusion model as an MTA model. To my knowledge network diffusion has not been used in the context of attribution before

.

References

Abhishek, Vibhanshu and Fader, Peter and Hosanagar, Kartik (2012), Media exposure through the funnel: A model of multi-stage attribution (August 17, 2012). Available at SSRN: http://ssrn.com/abstract=2158421 or http://dx.doi.org/10.2139/ssrn.2158421

Agee, T., Martin, B. (2001), Planned or Impulse Purchases? How to Create Effective Infomercials, Journal of Advertising Research, 41 (6), 35-42

Anderl, E., Becker, I., von Wangenheim, F., & Schumann, J. H. (2016), Mapping the customer journey: lessons learned from graph-based online attribution modeling. International Journal of Research in Marketing, 33(3), 457–474.

Bean, H. (1912), The Curve of Forgetting. In Repetition Effects over the Years. C. Craig and B. Sternthal, eds. New York: Garland Publishing Inc., reprinted from Archives of Psychology, 3, 21 (1912).

Berkowitz, D., Allaway, A., D'Souza, G. (2001), The Impact of Differential Lag Effects on the Allocation of Advertising Budgets across Media, Journal of Advertising Research, 41 (2), 27-36

Berman, R. R. (2015), Beyond the last touch: Aattribution in online advertising. Available at SSRN: http://ssrn.com/abstract=2384211 or http://dx.doi.org/10.2139/ssrn.2384211.

Bettman, J., Johnson, E., Payne, J. (1991), Consumer Decision Making "in" Bettman, J. (1991) Handbook of Consumer Behavior, Prentice-Hall, Engelwood Cliffs, 614.

Cacioppo, J., Petty, R. (1979), Effects of Message Repetition and Position on Cognitive Response, Recall, and Persuasion, Journal of Personality and Social Psychology, 37 (1), 97-109.

CNBC (2017 Uber just sued one of its ad agencies, and it points to growing mistrust with mobile advertising. Available at: https://www.cnbc.com/2017/09/19/uber-sues-fetch-for-ad-fraud.html, [26.4.2019]

Dalessandro, B., Stitelman, O., Perlich, C., & Provost, F. (2012), Causally motivated attribution for online advertising. NYU working paper series.

Engel, J., Blackwell, R. (1986), Consumer Behavior -5th edition , Chicago: Dryden Press, 633.

Hastie, T., Tibshirani, R., Friedman, J. (2009), The Elements if Statistical Learning: data Mining, Inference and prediction, 2nd edition, Springer, New York.

Iyer, A. V. (1999) "Modeling the Impact of Information on Inventories", in Tayur, S., Ganeshan, R. & Magazine, M. (eds.) Quantitative Models for Supply Chain Management, Kluwer, USA, pp. 337-358.

Joo, M., Wilbur, K., Zhu, Y. (2016) Effects of TV advertising on keyword search, International Journal of Research in Marketing, 33 (3), 508-523.

Kireyev, P., Pauwels, K., Gupta, S. (2015), Do display ads influence search? Attribution and dynamics in online advertising, International Journal of Research in Marketing, 33 (3), 475-490.

Kannan, P.K, Reinartz, W., Verhoef, P. (2016), The path to purchase and attribution modeling: Introduction to special section, International Journal of Research in Marketing 33 (3), 449-456

Li, H., Kannan, P. (2014) Attributing Conversions in a Multichannel Online Marketing Environment: An Empirical Model and a Field Experiment, Journal of Marketing Research, 51 (1), 40-56.

Lowell, N. (2013) The Curve: How Smart Companies Find High-Value Customers, Penguin, London, 243 s.

McKinsey (2015), Brand success in an era of Digital Darwinism. Available at: https://www.mckinsey.com/industries/high-tech/our-insights/brand-success-in-an-era-of-digital-darwinism, [25.9.2018]

Naik, P. (1999) Estimating the Half-life of Advertisements, Marketing letters, 10 (3), 351-362.

Naik, P., Mantrala, M., Sawyer, A. (1998), Planning Media Schedules in the Presence of Dynamic Advertising Quality, Marketing Science, 17 (3), 214-235.

NewZoo (2018), Mobile Revenues Account for More Than 50% of the Global Games Market as It Reaches $137.9 Billion in 2018. Available at: https://newzoo.com/insights/articles/global-games-market-reaches-137-9-billion-in-2018-mobile-games-take-half, [25.9.2018]

Olenski, S. (2017), Marketing's Next Big Hurdle: The Rising Cost Of Customer Acquisition, Available at: https://www.forbes.com/sites/steveolenski/2017/11/18/marketings-next-big-hurdle-the-rising-cost-of-customer-acquisition/#76e47e222298, [6.10.2018]

Peters, H. (2008), Game Theory a Multi-Leveled Approach, Springer-Verlag, Berlin, 362 p.

Shao, X., & Li, L. (2011), Data-driven multi-touch attribution models. KDD'11 (pp. 258–264).

Shapley, L. (1953), A Value for n-person Games. Contributions to The Theory of Games, Princeton University Press, Princeton, pp. 307-317.

Shocker, A., Ben-Akiva, M., Boccara, B., Nedungadi P. (1991), Consideration set influences on consumer decision-making and choice: Issues, models, and suggestions. Marketing letters 2(3), 181-197.

Takahashi, D. (2014), The rising cost of acquiring mobile-app users is hitting devs like a hurricane, Available at: https://venturebeat.com/2014/10/27/the-cost-of-acquiring-mobile-app-users-is-on-the-rise-again/, [6.10.2018]

Yadagiri M.M., Saini S.K., Sinha R. (2015) A Non-parametric Approach to the Multi-channel Attribution Problem. In: Wang J. et al. (eds) Web Information Systems Engineering – WISE 2015. WISE 2015. Lecture Notes in Computer Science, vol 9418. Springer, Cham, 338-352.

Xu, L., Duan, J., Whinston, A. (2014) Path to Purchase: A Mutually Exciting Point Process Model for Online Advertising and Conversion. Management Science 60(6):1392-1412. https://doi.org/10.1287/mnsc.2014.1952